BAB I
PENDAHULUAN

1.1 Latar Belakang

Twitter adalah sebuah situs jejaring sosial yang sedang berkembang pesat saat ini karena pengguna dapat berinteraksi dengan pengguna lainnya dari komputer ataupun perangkat *mobile* mereka dari manapun dan kapanpun. Setelah diluncurkan pada Juli 2006, jumlah pengguna Twitter meningkat sangat pesat. Pada September 2010, diperkirakan jumlah pengguna Twitter yang terdaftar sekitar 160 juta pengguna (Chiang, 2011).

Pengguna Twitter sendiri bisa terdiri dari berbagai macam kalangan yang para penggunanya ini dapat berinteraksi dengan teman, keluarga hingga rekan kerja. Twitter sebagai sebuah situs jejaring sosial memberikan akses kepada penggunanya untuk mengirimkan sebuah pesan singkat yang terdiri dari maksimal 140 karakter (disebut *tweet*). *Tweet* sendiri bisa terdiri dari pesan teks dan foto. Melalui *tweet* inilah pengguna Twitter dapat berinteraksi lebih dekat dengan pengguna Twitter lainnya dengan mengirimkan tentang apa yang sedang mereka pikirkan, apa yang sedang dilakukan, tentang kejadian yang baru saja terjadi, tentang berita terkini serta hal lainnya.

Pada tahun April 2010, jumlah *tweet* yang diposting mencapai 55 juta *tweet*/hari (Jackoway, dkk., 2011, hlm. 2), lalu kemudian pada tahun 2011, tercatat rata-rata sekitar 140 juta *tweet* telah dikirimkan oleh pengguna Twitter (Twitter Blog, 2011). Berbagai macam manfaat dapat diperoleh dari *tweet* dimulai dari *event detection* (deteksi kejadian, salah satunya bencana alam), prediksi pergerakan pasar saham, prediksi pemilu hingga penyebaran penyakit di suatu wilayah. Sebagai contoh, untuk prediksi pergerakan pasar saham, analisa dilakukan dengan cara menganalisa *tweet* yang berisi *mood* positif dan negatif

yang berkaitan dengan pasar saham seperti Dow Jones, S&P 500, NASDAQ (Zhang, dkk., 2010, hlm. 3). Contoh lainya yaitu *event detection*. Pada *event detection* (bencana alam), untuk memperoleh *tweet* yang akurat dan tepat sasaran, diterapkan semantik analisis *tweet* terhadap *keyword* yang muncul pada *tweet* (Sakaki, dkk., 2010, hlm. 2). Untuk mendapatkan manfaat dari *tweet* yang jumlahnya berlimpah ini, tentu saja dibutuhkan penelitian dan analisis terhadap *tweet* yang ada, salah satunya untuk penelitian *data mining* yang mempergunakan data dari *tweet.*

Data mining sendiri menurut Han dan Kamber (2006, hlm. 39) adalah sebuah upaya menemukan pola-pola yang menarik dari data yang berjumlah besar, yang dimana data-data tersebut bisa saja tersimpan di dalam *database*, *data warehouse*, ataupun di tempat penyimpanan lainnya. Begitu juga dengan data yang terdiri dari *tweet*, jumlah datanya berlimpah dan tentu saja memiliki pola-pola menarik yang bisa dimanfaatkan. Pada penelitian "*Twitter mood predicts the stock market*" (Bollen, dkk., 2010, hlm. 2), jumlah data *tweet* yang digunakan untuk menganalisis dan memprediksi pasar saham mencapai 9.853.498 *tweet.* Penelitian lain, "*Using prediction markets and twitters to predict swine flu pandemic*" (Ritterman, dkk., 2009, hlm. 3) menggunakan data *tweet* sebanyak 48 juta data *tweet.*

Untuk dapat mengumpulkan data *tweet* dalam jumlah yang besar tersebut, diperlukan sebuah sistem yang dapat mengumpulkan *tweet* yang tersedia sesuai dengan keyword tertentu. Dalam hal ini, Twitter sendiri telah menyediakan fasilitas Twitter API yang memberikan kemudahan untuk para peneliti untuk mengkoleksi dan mengumpulkan *tweet*. Twitter API memfasilitasi pengguna untuk dapat mengirimkan *request query* sebanyak 180 *request*/15 menit. Jika sebelum waktu 15 menit, *request* telah mencapai 180, maka harus menunggu 15 menit berikutnya untuk bisa melakukan *request* kembali.

Dalam mengumpulkan data *tweet*, peneliti ada yang menggunakan satu mesin dan juga beberapa mesin. Pada tahun 2009, ada penelitian "*What is Twitter, a Social Network or News Media?*" yang dilakukan oleh Kwak dkk. Penelitian ini

menggunakan 20 *whitelist* mesin dengan alamat IP yang berbeda (Kwak. dkk., 2009, hlm. 2) dan berhasil mengumpulkan 106 juta *tweet*. *Whitelist* adalah aturan yang ditetapkan Twitter (pada Twitter API v1) dengan memasukkan alamat mesin ke daftar putih Twitter dan memberikan keringanan dalam hal keterbatasan dalam melakukan permintaan dan saat ini (pada Twitter API v1.1) aturan *whitelist* sudah tidak berlaku.

Untuk menanggulangi keterbatasan permintaan pada Twitter API, dapat digunakan prinsip sistem terdistribusi. Prinsip sistem terdistribusi diterapkan agar di dalam satu waktu, proses pengumpulan *tweet* dapat dilakukan dengan lebih cepat dan mampu mengumpulkan lebih banyak *tweet* karena proses pengumpulan *tweet* akan didistribusikan ke beberapa *node*. Contohnya ingin mengumpulkan *tweet* dari beberapa keyword, maka sistem akan mendistribusikan beberapa keyword ini ke beberapa *node* yang tersedia. Dari beberapa *node* ini, jika terdapat *node* yang mati, maka tugasnya akan diambil alih oleh *node* lain sehingga sistem yang sedang dijalankan tidak mati. *Scalability* sistem juga harus diperhatikan. Sistem harus dapat menambahkan *node* secara dinamis jika diperlukan. Penelitian yang menerapkan sistem terdistribusi yaitu penelitian "*TwitterEcho – A Distributed Focused Crawler to Support Open Research with Twitter Data*" yang dilakukan oleh Bosnjak dkk. Penelitian ini menggunakan arsitektur distribusi terpusat. (Bosnjak, dkk., 2012, hlm. 2).

Penelitian Java dkk. (pada Bosnjak., dkk., 2012, hlm. 2) yang melakukan *crawl* Twitter dari tanggal 1 April hingga 30 Mei 2007 mendapatkan *tweet* sekitar 1,3 juta *tweet.* Jika dirata-ratakan, maka sekitar 15 *tweet* per menit yang bisa diperoleh. Dengan aplikasi *twitter crawler* yang akan dibuat, jika memaksimalkan parameter *count* dengan jumlah 100 dan distribusi menggunakan 3 mesin, maka bisa diperoleh *tweet* sebanyak 300 *tweet* dalam satu kali pencarian. Diharapkan dengan dikembangkannya aplikasi untuk pendistribusian proses pengumpulan *tweet* ini dapat mengoptimalkan pengumpulan data *tweet* dalam jumlah besar.

Pada skripsi ini, proses pendistribusian tugas pengumpulan *tweet* dilakukan menggunakan komputer virtual. Satu komputer virtual mewakili satu

node. Komputer virtual adalah representasi logis dari sebuah komputer di dalam perangkat lunak. (IBM Global Education White Paper., 2007, hlm. 3). Virtualisasi memungkinkan pengguna untuk menjalankan satu atau lebih mesin virtual secara bersamaan, yang masing-masing mesin virtual memiliki sistem operasinya di atas komputer fisik tunggal (Li., 2010, hlm. 12).

Skripsi ini membahas tentang pendistribusian proses pengumpulan *tweet* sehingga mampu mengumpulkan data *tweet* dalam jumlah besar yang bermanfaat untuk peneliti lainnya.

1.2 Rumusan Masalah

Berdasarkan latar belakang masalah yang ada, maka permasalahan dalam skripsi ini dirumuskan sebagai berikut:

1. Bagaimana melakukan pendistribusian proses pengumpulan *tweet* ke banyak *node* sehingga akan bisa diperolah *tweet* yang lebih banyak dalam satu waktu?
2. Bagaimana mengembangkan *prototype* sistem untuk memudahkan pengumpulan data *tweet* dalam jumlah besar?

1.3 Batasan Masalah

Untuk memfokuskan penelitian, ada beberapa batasan masalah, yaitu sebagai berikut:

1. Twitter API yang digunakan pada penelitian adalah Search API yang merupakan bagian dari REST API.
2. Data yang diambil dari *twitter* hanyalah sebatas data *tweets* yang sesuai dengan *keyword* yang dimasukkan oleh pengguna.
3. Proses pendistribusian yang dilakukan menggunakan komputer virtual.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dikemukakan, maka tujuan dari penelitian yang akan dilakukan ini adalah:

1. Dapat mendistribusikan proses pengumpulan *tweet* ke beberapa *node* yang tersedia sehingga dapat diperoleh jumlah *tweet* yang besar dalam waktu yang lebih cepat dibandingkan dengan pengambilan dengan satu *node*.
2. Dapat mengembangkan *prototype* sistem yang dapat memudahkan pengumpulan tweet dalam jumlah besar.

1.5 Manfaat Penelitian

Manfaat yang ingin diperoleh dari penelitian ini adalah:

1. Mempermudah serta mempersingkat waktu proses pengumpulan *tweet.*
2. Menambah wawasan serta dapat menerapkan ilmu yang diperoleh di perkuliahan.
3. Dapat menjadi bahan rujukan bagi peneliti lain dalam penelitiannya yang memiliki keterkaitan dengan penelitian ini.

1.6 Sistematika Penulisan

Sistematika penulisan skripsi ini adalah sebagai berikut:

BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Bab ini berisi tentang teori-teori serta konsep-konsep yang berfungsi sebagai sumber atau alat dalam memahami yang akan diterapkan pada penelitian yang akan dilakukan.

BAB III METODOLOGI PENELITIAN

Bab ini berisi tentang penjelasan tahap-tahap yang akan dilakukan dalam penelitian, serta hasil penelitian, dan pembahasan dari hasil penelitian.

BAB IV HASIL PENELITIAN DAN PEMBAHASAN

Bab ini berisi pemaparan hasil penelitian pendistribusian proses pengumpulan tweet disertai fakta yang diperoleh selama penelitian.

BAB V KESIMPULAN DAN SARAN

Berisi kesimpulan yang dapat diambil dari penelitian dan saran untuk pengembangan penelitian selanjutnya.